Root Cause Recognition

Feasibility and Technology Selection Phase

Root Cause Recognition

- Objective: Classify the state all of customer equipment instances at any specific point in time. Labels include: typical, atypical/unknown, and root cause *x*
  - RCR are time series based on time series
  - Time series is seen as collection of “characteristics”,
  - RCR labels evolve over time
  - Key assumption: equipment instances’ characteristic time series contains sufficient information resolution to resolve root cause
    - More specifically: given a continuous TS of equipment instance characteristics, it is possible to identify pattern sequences of various length that correspond to root causes

Demystifying Machine Learning

- No magic, just patterns
  - We are looking for patterns of characteristics that match (to some extent) the patterns that we are looking for (classification) or patterns that are different from enough typical (anomalies)
    - If patterns are conserved (repeated) they represent important behaviors
    - The challenge is in matching patterns which are never exact matches efficiently

Root Cause Recognition is Worse Case ML

- Time series are special
  - Adds additional dimension, time, which warps the patterns
  - Time series are windowed. Different windows enable different patterns to be identified. So the same time series can be viewed with different “lenses” and have different results
  - The data set constantly evolves (forensic analysis aside) so patterns constantly evolve.
- Input vectors are multivariate
  - Multivariate introduce large dimensional spaces (and associated computation overhead)
- Root Cause Recognition is classification (labeling) of multivariate time series data
  - I haven’t found published work of projects at this scale.

Classification Approaches

- Supervised Learning
  - The learning algorithm is presented with example inputs (characteristics) and their desired outputs (labels), and the goal is to create general rules that maps characteristics to labels
- Unsupervised Leaning
  - No labels are given to the learning algorithm, leaving it on its own to find structure in the input characteristics. Can also be used to identify candidate labels
- Semi-supervised Learning
  - There are some labels, but they are noisy, limited or imprecise.

Some Common Classification Techniques

- Distance based (KNN with DTW)
- Interval based (TimeSeriesForest)
- Dictionary based (BOSS, cBOSS)
- Frequency based (RISE)
- Shapelet based (Shapelet Transform Classifier)
- Neural Networks

Factors Considered to Select Approach

- Efficiency
- Appropriateness to streaming
- Hyper-parameter tuning
- Training Data (labels)

Approach Chosen to Pursue

- Machine augmented, semi-supervised leaning approach
  - A dictionaries of root causes will be maintained by domain experts who label discord patterns found in time series data.
  - Labelling is an ongoing process. Labels are assumed to be noisy and incomplete
- Root cause recognition will be done using a multi-variant version of the SDTS (Scalable Dictionary learning for Time Series) algorithm [https://www.cs.ucr.edu/~eamonn/WeaklyLabeledTimeSeries.pdf].
  - Shapelet dictionary-based approach
  - Multivariate variation is unique
- Dictionary maintenance will be done using a variation of the LBLR (Like-Behavior Labeling Routine) described in https://www.cs.ucr.edu/~eamonn/Efficient%20and%20Effective%20Labeling%20of%20Massive%20Entomological%20Datasets.pdf
  - Finding RCR candidate shapelets is an unsupervised data mining task (KNN-like)
  - Fusion point charter will be modified to provide domain experts visualization of the found shapelets

Why this Approach

- Underlying technology is Matrix Profiles
  - Very fast and efficient
    - Can process streaming data in real-time (faster than 15-minute for large number of equipment instances with multi-variant time series)
  - Does well with noisy data.
  - Identifies Discords (anomalies) and Motifs with the same calculation
  - Well supported by community and open source software

Challenge Thus Far

- Dirty, noisy data
- Number of characteristics
- Multi-variant time series machine learning is the cutting edge of ML
    - Attempted many approaches, some where promising but impossible to implement at Resolute scale.
- Personal learning curve.

Let’s look at some code…

- Single Variable Exploration
- Multivariate Motif Exploration
- Multivariate Discord Exploration

Conclusions & Next Steps

- Discord discovery using multivariate matrix profiles is a feasible approach.
  - Finds discords and motifs in AHU data
  - Multivariate MP is definitely performant
  - Work still needs to be done to find the most significant “repeating discords” (perhaps “rare motif” is a better classification)
- Next Steps
  - Challenge of problematic data still exists
  - Develop canonical list of motifs for an AHU instance and compare it against domain expert developed list

Longer term

- Next phase building the prototype
  - Root cause labeling utility
    - Scans input vectors looking for unlabeled discords. Results are presented to domain expert in Fusion point charter-like format for labeling.
  - Root cause classification process
    - Periodically labels all equipment instances with their current state